ICT for Health Science Research
A. Shabo (Shvo) et al. (Eds.)

A Concept for a Data Dictionary System Supporting for Clinical Research

Moritz STRICKLER[a,1], Chantal ZBINDEN[a,1], Ramon SACCILOTTO[b] and Kerstin DENECKE[a]
[a] *Bern University of Applied Sciences, Biel, Switzerland*
[b] *University of Basel, Basel, Switzerland*

Abstract. Clinical trials form the backbone of evidence-based medicine. They are indispensable tools that make it possible to compare medical interventions or test medical devices for their efficacy and safety. Each study design exactly defines which data is to be collected. Case report forms (CRFs) are used to document the collected data. Preparing CRFs is complicating, time consuming and requires extensive knowledge in the fields of medicine, data management and statistics. In addition, there are no global standards for CRF design, which means that each research institution produces CRF data definitions at its own discretion. This hampers the exchange of data definitions among different research groups and variants of CRFs might be created for a similar study design. To address these problems, we developed a concept for a freely accessible portal in the form of a web application in which definitions for CRFs, variables and tables can be created. The created data definitions can be exported from the portal to be transferred to common electronic data capture systems (EDC) that can then generate CRFs using the definition. The overall objective of the project is to develop a data dictionary system that is used during the entire workflow of a study and that enables sharing and re-use of metadata.

Keywords. Metadata, Case Report Form, Case Report Form Design, Standard Templates, Data Dictionary, Electronic Data Capture System, Clinical Research

1. Introduction

Case report forms (CRFs) [1] are questionnaires tailored to a specific study design in which the necessary examination data of patients are captured. The collected data are usually recorded in a coded form. The preparation of a CRF for a specific research project causes huge efforts for a research group and can produce serious implications when definitions are not well specified. It must be ensured that different persons such as study nurses record the same information when completing a CRF (e.g. in the same format and measurement unit). Therefore, the data definitions and specifications of CRFs should enable to collect data in sufficient detail without ambiguity, unnecessary details and should avoid redundancy [2].

For clinical trials, usually electronic data capture (EDC) systems are used, which already have a user interface to define CRFs. However, they are usually closely coupled to the software provider and often too complicated for the average user to handle. This

[1] Both authors share the first authorship and contributed equally to this work, corresponding authors, Moritz Strickler, Chantal Zbinden, Bern University of Applied Sciences, Quellgasse 21, 2501 Biel, Switzerland; E-mail: moritzkaspar.strickler@students.bfh.ch, chantal.zbinden@students.bfh.ch.

leads to the fact that data definitions and thoughts about analytic steps are often not transacted and thereby neglected with the definition of CRFs. As a result, problems in the statistical evaluation of the forms or during the setup of the study database can occur [2]. Due to lack of generally accepted standards for data capturing and data definition in CRFs, sharing and reusing of clinical data is complicated to put into practice. Furthermore, this is normally impossible due to data privacy policy.

To address these limitations, we introduce a concept for managing and sharing data definitions in the form of data dictionaries through a user-friendly web application in which data dictionaries can be created. A data dictionary is a centralized repository that provides metadata about specific data, i.e. meaning, relationships to other data, origin, usage, and format. Metadata in the form of data dictionaries can be made accessible without any problems and offer the potential to improve the reuse and harmonization of data across projects [3]. Accessible to researchers on a Swiss national level, our system enables developing, administering and sharing of comprehensive, interoperable data dictionaries for clinical research projects.

2. Material and methods

The time frame for the project is one year with implementation start in January 2019. The requirements for the portal were primarily determined through an interview with the project partner and stakeholder in a university hospital. In addition, further requirements were identified through literature search and the analysis of existing solutions for the definition of CRFs. The following factors were taken into account when developing the concept: 1) fulfilment of the requirements identified, 2) compliance with the project duration of one year, 3) compliance with project and operating costs. After the development of the portal, a test series is planned. The portal will be tested together with the EDC system SecuTrial®.

Currently, there are no universal CRF design standards, however, there are conventions and some 'best' practices do. The Clinical Data Interchange Standards Consortium (CDISC), which focuses primarily on regulated studies, has proposed such standards. CDISC defined the Operational Data Model (ODM), which is an international open standard for metadata and data in clinical trials as an XML format. This standard is supported by many EDC systems such as SecuTrial®. In addition, CDISC ODM can be semantically annotated. For these reasons, CDISC ODM has been chosen in our concept for exporting data definitions.

3. Results

3.1. Requirements

The requirements are separated by user roles. They include researchers, study coordinators, database managers, data entry staff, statisticians and clinical trial units (CTU) as regulators. Typically, researchers (Principal Investigator, PI) are familiar with standard office tools; data managers work with EDC systems and statisticians use specific statistical software. Therefore, a typical study consists normally of three different data models because each user group specifies its own [4].

Researchers should be able to use the portal to design CRFs in a simple and collaborative way at the beginning of a clinical study, using a large pool of existing data definitions from other studies or standard templates. A commenting and reference function is intended to provide improved cooperation in CRF development. The portal's import and export modules should enable synchronization with EDC systems commonly used in Switzerland and should be uploaded as a PDF export to the Business Administration System for Ethics Committees (BASEC) for study submission. The data catalog is intended to be compatible with the common classifications of medicine, e.g. SNOMED CT or LOINC. In addition, according to the FAIR Data Principles, the metadata definitions have to be made available to other researchers after completion of a study.

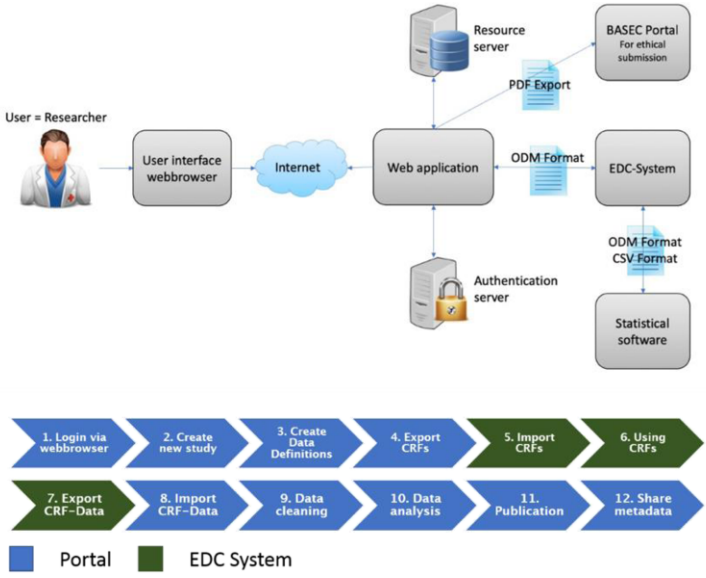

Figure 1. System architecture and process

3.2. Concept

The system architecture of the portal is shown in Figure 1. Defined CRFs and related data specification will be stored on a central data store (resource server) of the CTU of the collaborating hospital. Access to the server and correspondingly to the data dictionary system will be provided as Software-as-a-Service (SaaS). In this way, hospitals or research institutions will not have to run and maintain the system on their places. The system will not communicate directly with other systems such as EDC systems, but will offer import and export functions via XML files based on the CDISC ODM. In this way, the CRF could be exported as PDF and will be able to be transmitted to the BASEC portal for approval. The user administration will be realized via an authentication server. CDISC ODM is suitable for the structure representation of all relevant data and information about the process (including visits and follow-ups) of a clinical trial as well as the required metadata including versioning of a study [3,5].
There will be two different types of users for our system: the researchers who can create data definitions and share data through the application and the CTU of the collaborating

hospital that interacts as administrator of the portal. The access to the portal by researchers of any hospital is independent from the maintaining CTU.

The front-end system will visualize the researcher's entered metadata of studies and a table-based representation of the data definitions. The researcher has the option to assemble data definitions by himself by using various data definition building blocks ordered by subject areas, such as inclusion or exclusion criteria. They are able to search the templates and data definition blocks by topic and the portal will also suggest other studies using the same building blocks.

In addition, variables and blocks of variables can be selected and integrated into a new collection or a data definition. Created data definition collections can be released to the CTU of the collaborating hospital for checking for correctness. The metadata for the created forms can optionally be made available to other research groups. The data definitions can be exported in ODM format from the portal and imported for use in the standard EDC systems.

4. Discussion

Our data dictionary portal will enable the web-based creation of data definitions that span the whole lifecycle of a clinical study. It will support users during their data definition processes by providing logical rules and thus helps to avoid common errors. An existing portal for medical data models (MDM) has also addressed this problem. The MDM-portal is a German and European open-access metadata-repository initiated for scientific purposes that supports a user in generating, analyzing, releasing and reusing medical forms [6]. The functions of the MDM-portal are far too limited for everyday use. In particular, the data has to pass through several phases in a study which is not mapped into the system. First, the data is entered into the CRF. Then, the data is usually exported in tabular form by the IT system (usually enriched with data that does not necessarily have to be entered in the CRF) and finally the data is prepared by the statisticians and made available as final tables. Our data dictionary system will be able to map all steps and thereby allow the "life cycle" of a variable to be tracked over the entire process (i.e. from which CRF does the variable in my final table originate). Another important missing function in the MDM-portal is the time dependency between the forms. Studies often run in relatively fixed visit plans and the timing and dependencies between the forms is often important to understand and interpret the context of the data correctly. Further, groups of fields that can be repeated together - e.g. for a medication these would be substance, quantity, date, person administering - are not really included in the MDM definitions. Our system will offer a guided tour on creating forms and will remind the user, if he has made changes to a data definition, to have the corresponding CRF approved by the ethics committee. Additionally, it will inform whether other studies use the same data models. However, we believe that MDM can benefit from our data dictionary system which could ideally automatically transfer in future the data definitions to the MDM register. Even though developed specifically for Switzerland, our concept for the portal can be easily transferred to a different country.

The use of our portal will cause a significant change in the work process of the participants. This could lead to acceptance problems and must be taken into account in the test series. According to Löbe et al. the quality of the content is more relevant to users than the efficiency of the documentation [7]. Other non-technical obstacles could be the intellectual property. The MDM portal has tried to solve this with licenses.

Another project is the Clinical Data Acquisition Standards Harmonization (CDASH) of CDISC. CDASH addresses data collection standards through standardized CRFs [8]. CDISC uses also the ISO/IEC 11179 design for CSHARE. The ISO/IEC 11179 standard provides a metadata registration as a database of metadata, which supports the functionality of the registry. The registration achieves three main goals: identification, origin and quality control. CSHARE is a repository for domain-specific research questions and answer sets [9]. A limitation of CDASH for realizing a metadata repository is that some specifications require data-entry staff to use calculators, instead of programming computations directly into electronic CRFs [10]. Leroux et al. show that FHIR can semantically enrich the ODM data. By exploiting the rich information model in FHIR, clinical data can be organized in a manner that preserves its organization, but captures its context [11]. A further approach is taken by the openEHR archetypes as a basis for HL7 Clinical Document Architecture templates are agreed-upon specifications that support computable definitions of clinical concepts [12]. Since it has proven that adopting the openEHR approach is highly desirable to multi-center clinical trials [13], it would be interesting for future research to consider openEHR also in the portal.

References

[1] ICH Guidance E6: Good Clinical Practice, P. 9.: US HHS, US FDA, CDER, CBER, 1996. Available from: https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R1_Guideline.pdf [Last accessed on 2018 Oct 28].

[2] S. Bellary, B. Krishnankutty, M.S. Latha. Basics of case report form designing in clinical research. Perspect Clin Res. Oktober 2014;5(4):159–66.

[3] V. Huser, C. Sastry, M. Breymaier, A. Idriss, J.J. Cimino. Standardizing data exchange for clinical research protocols and case report forms: An assessment of the suitability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM). J Biomed Inform. October 2015; 57:88–99.

[4] M. Dugas. Integrated Data Management for Clinical Studies: Automatic Transformation of Data Models with Semantic Annotations for Principal Investigators, Data Managers and Statisticians. PLOS ONE 2014; Volume 9, Issue 2, e90492

[5] S. Hume, J. Aerts, S. Sarnikar, V. Huser. Current applications and future directions for the CDISC Operational Data Model standard: A methodological review. Journal of Biomedical Informatics. 1. April 2016;60:352–62.

[6] M. Dugas. Design of case report forms based on a public metadata registry: re-use of data elements to improve compatibility of data. Trials. 2016; 17: 566.

[7] M. Löbe. User Expectations of Metadata Repositories for Clinical Research. Stud Health Technol Inform. 2018;253:60-4.

[8] CDISC, Clinical Data Acquisition Standards Harmonization: Basic Data Collection Fields for Case Report Forms. Draft version 1.0. http://www.cdisc.org/cdash [Last accessed on 2018 Oct 28].

[9] S. Ngouongo, M. Löbe, J. Stausberg. The ISO/IEC 11179 norm for metadata registries: Does it cover healthcare standards in empirical research? Journal of Biomedical Informatics, Volume 46, Issue 2, 2013, Pages 318-327, ISSN 1532-0464

[10] A. Roehrs, C.A. da Costa, R.D.R. Righi, S.J. Rigo, M. Wichman. Toward a Model for Personal Health Records Interoperability. IEEE J Biomed Health Inform. 2018 May 14. doi: 10.1109/JBHI.2018.2836138.

[11] H. Leroux, A. Metke-Jimenez, M.J. Lawley. Towards achieving semantic interoperability of clinical study data with FHIR. J Biomed Semantics. 2017;8(1):4.

[12] H. Leslie. International developments in openEHR archetypes and templates. HIM J 2008; 37:38-9

[13] S. Garde, P. Knaup, T. Schuler, E. Hovenga. Can openEHR Archetypes Empower Multi-Centre Clinical Research? Stud Health Technol Inform. 2005;116:971-6